

Svar til 27611 sommereksamen 2014:

ER proteiner, KDEL motiv og KDEL receptor

Opgave 1 - Karakterisering af KDEL receptoren

Spørgsmål a: Der er **942** proteiner i UniProt, der hedder "ER lumen protein retaining receptor"; **33** i Swiss-Prot (reviewed) og **909** i TrEMBL (unreviewed).

Søgestreng: `name:"ER lumen protein retaining receptor"`

Spørgsmål b: Der er **3** menneskelige hits i Swiss-Prot (reviewed).

Søgestreng: `name:"ER lumen protein retaining receptor" AND organism:"Human [9606]"`

Spørgsmål c: "ER lumen protein retaining receptor 1" fra menneske har accession-kode **P24390** og UniProt ID **ERD21_HUMAN**. Gen-navnet er **KDEL R1** eller **ERD2.1**.

Spørgsmål d: **Ja**, ERD21_HUMAN er et transmembranprotein. Det kan ses under **Subcellular location**, hvor der står "Multi-pass membrane protein", under **Keywords**, hvor der står "Transmembrane" samt i **feature-tabellen**, hvor der er flere "Transmembrane" features. Der er **7** transmembran-segmenter, men deres positioner er **ikke eksperimentelt bestemt** (der står "Potential" ud for dem alle).

Spørgsmål e: "ER lumen protein retaining receptor" fra bagegær (*Saccharomyces cerevisiae*) i Swiss-Prot har accession-kode **P18414** og UniProt ID **ERD2_YEAST**. Gen-navnet er **ERD2**. Man finder (kun) den med søgestrengen `(name:"ER lumen protein retaining receptor" AND organism:"Saccharomyces cerevisiae") AND reviewed:yes`

Spørgsmål f: **Ja**, der er en forskel. Om ERD21_HUMAN står der "This receptor recognizes the C-terminal K-D-E-L motif". Om ERD2_YEAST står der "This receptor strongly recognizes H-D-E-L and weakly recognizes D-D-E-L and K-D-E-L". Gær foretrækker altså H på første position af motivet, hvor mennesker foretrækker K.

Spørgsmål g: Her er alignmentet:

ERD21_HUMAN	1	MNLFREFLGDLSHLLAIILLLLKIWKSRSCAGISGKSQVLFAVVFTARYLD	50
		. .	
ERD2_YEAST	1	MNPFRLIGDLSHLTSILILIHNIKTTRYIEGISFKTQTLVALVFITRYLD	50
ERD21_HUMAN	51	LFT-NYISLYNTCMKVVI-ACSFTTVWLIYSKFKATYDGN----HDTFR	94
		. :	

Opgave 2 - Karakterisering af KDEL motivet fra dyr og svampe

Spørgsmål a: 35607

Søgestreng: `annotation:(type:location "endoplasmic reticulum")`

Spørgsmål b: 6642

Søgestreng: `annotation:(type:location "endoplasmic reticulum" confidence:experimental)`

Spørgsmål c: 6146

Søgestreng: `annotation:(type:location "endoplasmic reticulum lumen")`

Spørgsmål d: 1776

Søgestreng: `annotation:(type:location "endoplasmic reticulum") AND annotation:(type:signal)`

Spørgsmål e: 19016

Søgestreng: `annotation:(type:location "endoplasmic reticulum") AND fragment:no`

Spørgsmål f: 519

Søgestreng: `annotation:(type:location "endoplasmic reticulum lumen") AND annotation:(type:signal) AND fragment:no`

Spørgsmål g:

Menneske (*Homo sapiens*) tilhører riget **Metazoa** (dyr).

Bagegær (*Saccharomyces cerevisiae*) tilhører riget **Fungi** (svampe)

Spørgsmål h:

Metazoa: **307**

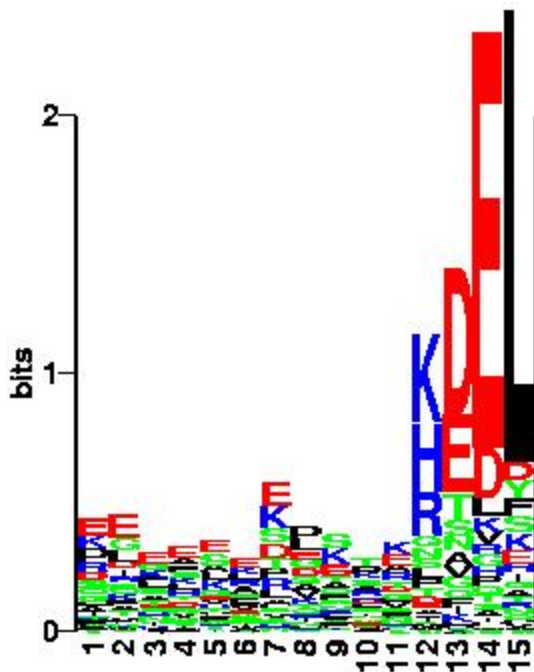
Søgestreng: `annotation:(type:location "endoplasmic reticulum lumen") AND annotation:(type:signal) AND fragment:no AND taxonomy:Metazoa`

Fungi: **56**

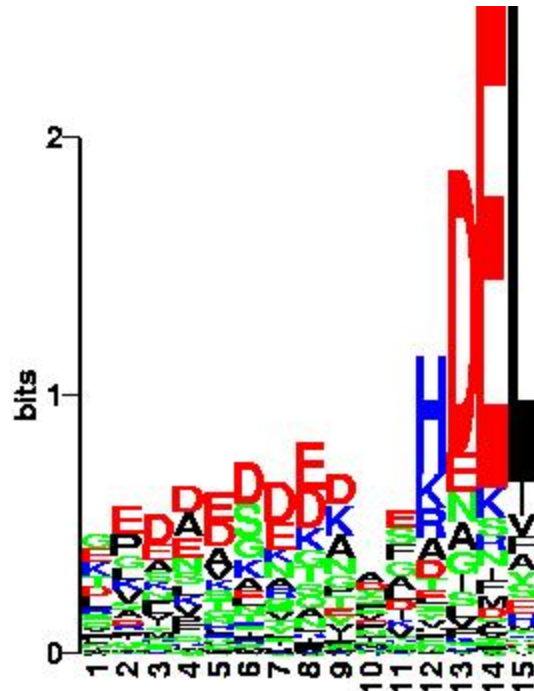
Søgestreng: `annotation:(type:location "endoplasmic reticulum lumen") AND annotation:(type:signal) AND fragment:no AND taxonomy:Fungi`

Spørgsmål i:

Metazoa:



Fungi:



Positionerne 14 og 15 indeholder mest information i begge logoer. Man kan direkte aflæse konsensus-sekvenserne for KDEL-motiverne ved at tage de øverste bogstaver i de sidste fire positioner (12-15): KDEL for Metazoa og HDEL for Fungi. Forskellen mellem de to riger er tydeligst i position 12, hvor H er meget mere sandsynlig end K for Fungi, men man kan også se en forskel i position 13, D er meget mere sandsynlig end E for Fungi, mens den kun er lidt mere sandsynlig end E for Metazoa. Der er også en forskel i positionerne 1-9, hvor der er en vis præference for D og E i Fungi, men kun en ganske svag præference for E i Metazoa.

Opgave 3 - Fylogeni af KDEL receptoren

Spørgsmål a: Hvis man selv vil downloade sekvenserne i FASTA format fra UniProt, skal man bruge følgende søgestreng:

name:"ER lumen protein retaining receptor" AND reviewed:yes

og derefter klikke på den orange "Download" knap og vælge "Download" linket under "FASTA". Men det er helt OK blot at bruge den fil, der blev opgivet.

Efter de beskrevne rettelser ser FASTA filen således ud:

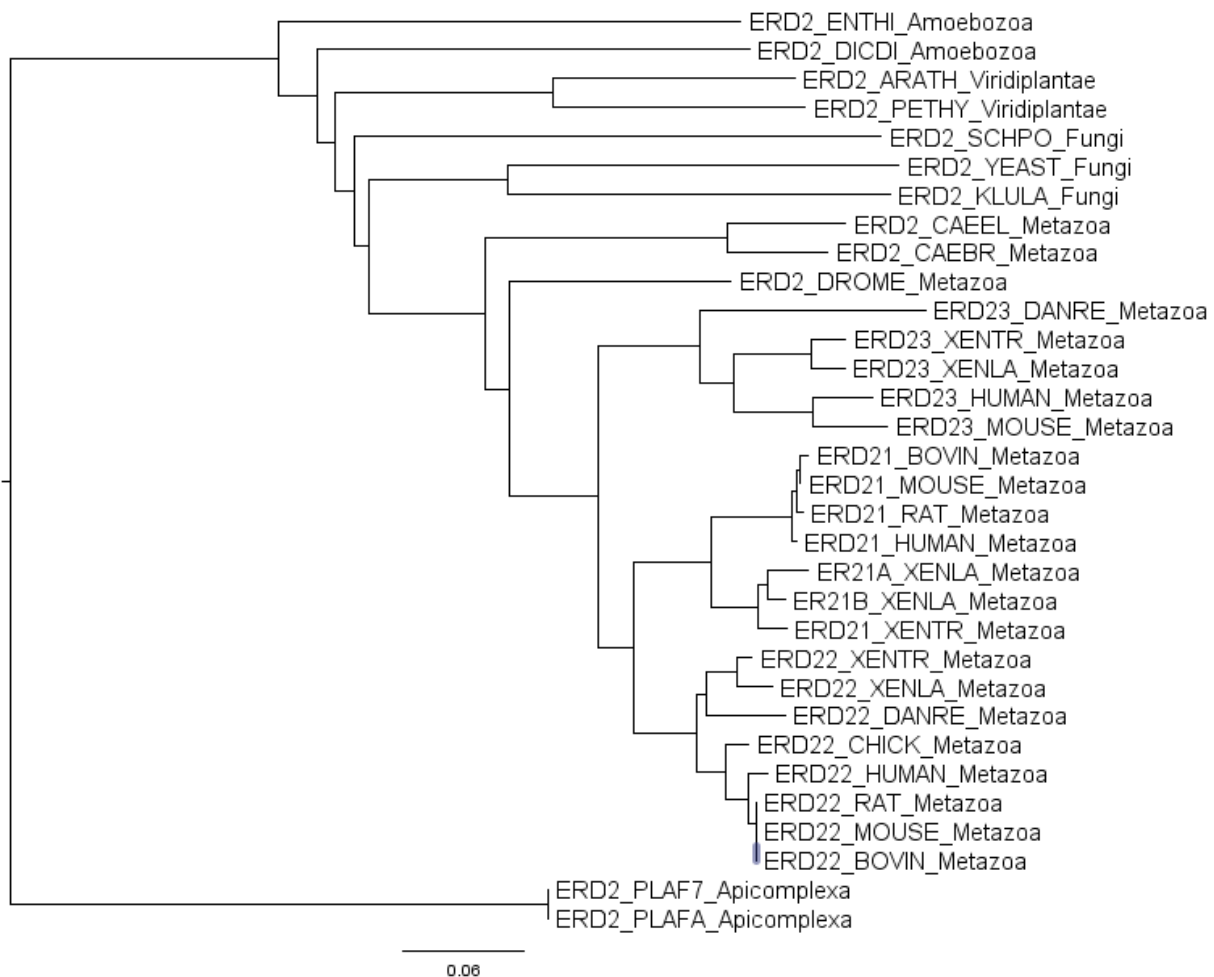
```
>ER21A_XENLA_Metazoa
MNI FRFLGDISHL SAI FILL LK IWKSRSCAGISGKSQ L LFAIVFTARYLDLFTNYISFY N
TSMKVVYVASSYATVWMIYSKFKATYDGNHDTFRVEFLIVPTAILAFLVNHDFTPLEIFW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGIYRTLYLFNWIWRYQFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGK KLSLPA
>ER21B_XENLA_Metazoa
MNI FRFLGDISHL SAI I I L L L K IWKSRSCAGISGKSQ L LFAIVFTTRYLDLFTNFISFY N
TSMKVVYVASSYATVWMIYSKFKATYDGNHDTFRVEFLIVPTAILSFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGIYRTLYLFNWIWRYQFEEFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGK KLSLPA
>ERD21_BOVIN_Metazoa
MNI FRFLGDL SHLLAI I L L L L K IWKSRSCAGISGKSQ V LFAVFTARYLDLFTNYISLY N
TCMKVVYIACSFTTVWMIYSKFKATYDGNHDTFRVEFLVIPTAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGK KLSLPA
>ERD21_HUMAN_Metazoa
MNI FRFLGDL SHLLAI I L L L L K IWKSRSCAGISGKSQ V LFAVFTARYLDLFTNYISLY N
TCMKVVYIACSFTTVWLIYSKFKATYDGNHDTFRVEFLVVP TAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGK KLSLPA
>ERD21_MOUSE_Metazoa
MNI FRFLGDL SHLLAI I L L L L K IWKSRSCAGISGKSQ V LFAVFTARYLDLFTNYISLY N
TCMKVVYIACSFTTVWMIYSKFKATYDGNHDTFRVEFLVVP TAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGK KLSLPA
>ERD21_RAT_Metazoa
MNI FRFLGDL SHLLAI I L L L L K IWKSRSCAGISGKSQ V LFAVFTARYLDLFTNYISLY N
TCMKVVYIACSFTTVWMIYSKFKATYDGNHDTFRVEFLVVP TAVLAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGK KLSLPA
>ERD21_XENTR_Metazoa
MNI FRFLGDISHL SAI I L L L L K IWKSRSCAGISGKSQ L LFAIVFTTRYLDLFTNFISLY N
TSMKVVYVASSYATI WMIYSKFKATYDGNHDTFRVEFLIVPTAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGIYRALYLFNWIWRYQFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGK KLSLPA
>ERD22_BOVIN_Metazoa
MNI FRLTGDL SHLAAI V I L L L K IWKTRSCAGISGKSQ L LFAVFTTRYLDLFTSFISLY N
TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHD FSPLEILW
TFSIYLESVAILPQLFMISK TGEAETITTHYLFFLGLYRALYLVNWIWR FYFEGFFDLIA
VVAGVVQTILY CDDFFLYITKVLKGK KLSLPA
>ERD22_CHICK_Metazoa
MNI FRLTGDL SHLAAI I I L L L K IWKSRSCAGISGKSQ L LFAVFTTRYLDLFTSFISLY N
TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLIVPVGGLSFLVNHD FSPLEILW
TFSIYLESVAILPQLFMISK TGEAETITTHYLFFLGLYRALYLVNWIWR YFYFEGFFDLIA
VVAGVVQTVLY CDDFFLYITKVLKGK KLSLPA
>ERD22_HUMAN_Metazoa
```

MNIFRLTGDLSHLAAIVILLKKIWKTRSCAGISGKSQQLFALVFTTRYLDLFTSFISLYN
 TSMKVIYLACSYATVYLIYKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRALLYLNWIWRIFYFEGFFDLIA
 VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
 >ERD22_DANRE_Metazoa
 MNIFRLTGDLSHLAAIIILLKKIWKSRSCAGISGKSQILFALVFTTRYLDLLTSFISLYN
 TCMKVIYIGCAYATVYLIYAKFRATYDGNHDTFRAEFLVVPVGGLAFLVNHDFSPLLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFCLGVYRALLYLNWIWRIFYFEGFFDMIA
 IVAGVVQTILYCDFFLYVTKVLKGKKLSLPA
 >ERD22_MOUSE_Metazoa
 MNIFRLTGDLSHLAAIVILLKKIWKTRSCAGISGKSQQLFALVFTTRYLDLFTSFISLYN
 TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRALLYLNWIWRIFYFEGFFDLIA
 VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
 >ERD22_RAT_Metazoa
 MNIFRLTGDLSHLAAIVILLKKIWKTRSCAGISGKSQQLFALVFTTRYLDLFTSFISLYN
 TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRALLYLNWIWRIFYFEGFFDLIA
 VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
 >ERD22_XENLA_Metazoa
 MNVFRLSGDLCHLAAIIILLKKIWNRSRSCAGISGKSQQLFAMVFTTRYLDLFTSFISLYN
 TSMKVIYMGAYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSVLVNHDFSPLLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRALLYLNWIWRFSFEGFFDLIA
 IVAGVVQTILYCDFFLYVTKVLKGKKLSLPA
 >ERD22_XENTR_Metazoa
 MNVFRLSGDLSHLAAIIILLKKIWKSRSCAGISGKSQQLFALVFTTRYLDLLTSFISLYN
 TSMKVIYIGCAYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSVLVNHDFSPLLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRALLYLNWIWRYSFEGFFDLIA
 IVAGVVQTILYCDFFLYVTKVLKGKKLSLPA
 >ERD23_DANRE_Metazoa
 MNIFRLSGDVCHLIAIIILLFKIWRKSKSCAGISGKSQVLFALVFTTRYLDLFTSYISAYN
 TVMKVVYLLLAYSTVGLIFFRNSYDSESDSFRVEFLVVPVAGLSFLENYAFTPLEILW
 TFSIYLESVAILPQLFMITKTEGAEITAHYLLFLGLYRALLYLANWLWRFHTEGFYDQIA
 VVSGVVQTIFYCDFFLYFTRVLRGSGKMSLPMPV
 >ERD23_HUMAN_Metazoa
 MNVFRILGDLSHLLAMILLKGIWRKSKCKGISGKSQILFALVFTTRYLDLFTNFISIYN
 TVMKVVFLLCAYVTVYMIYKFRKTFDSENDTFRLEFLVVPVIGLSFLENYSFTLLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRALLYLANWIWRYQTENFYDQIA
 VVSGVVQTIFYCDFFLYVTKVLKGKKLSLPMPI
 >ERD23_MOUSE_Metazoa
 MNVFRILGDLSHLLAMILLVKIWRKSKSCAGISGKSQILFALVFTTRYLDLFSNFISIYN
 TVMKVVFLLCAYVTVYMIYKFRKTFDIENDTFRLEFLVVPVIGLSFLENYSYTPMEVLW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRLLYLANWIWRYQTENFYDQIS
 VVSGVVQTIFYCDFFLYVTKVLKGKKLSLPVPV
 >ERD23_XENLA_Metazoa
 MNIFRILGDIVSHLAAIIILLKMWKSKSCAGISGKSQQLFALVFTTRYLDLFTVFISPYN
 TVMKIIFLACAYVTVYLIYKLRKSYDSENDTFRLEFLVVPVIGLSFLENYEFTPLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRVLYLANWIWRYHTEKFYDQIA
 VVSGVVQTIFYFDFFLYVTKVLKGKKLSLPMVPV
 >ERD23_XENTR_Metazoa
 MNVFRISGDIVSHLAAIIILLKMWKSKSCAGISGKSQQLFALVFTTRYLDLFTVFISAYN
 TVMKIVFLVCAYVTVYLIYKFRKAYDSENDTFRLEFLVVPVIGLSFLENYEFTPLEILW
 TFSIYLESVAILPQLFMISKTEGAEITITTHYLFGLYRVLYLANWIWRYHTEKFYDQIA
 VVSGVVQTIFYFDFFLYITKVLKGKKLSLPMVPV
 >ERD2_ARATH_Viridiplantae
 MNIFRFAGDMSHLISVLILLKKIYATKSCAGISLKTQELYALVFLTRYLDLFTDYVSLYN
 SIMKIVFIASSLAIVWCMRRHPLVRRSYDKDLDTFRHQYVVLACFVLGLILNEKFTVQEV

FWAFSIYLEAVAILPQLVLLQSRGNDNLTGQYVVFLLGAYRGLYIINWIYRYFTEDHFTR
WIACVSGLVQTALYADFFYYYYISWKTNTKLKLP
>ERD2_CAEER_Metazoa
MNI FRITADLAHVAIAIVILLKKIWKSRSCGISGRSQILFAVTFTRYLDLFTSFYSLYN
TVMKVLFLAGSIGTVYLMWVKFKATYDRNNDTFRIEFLVIPSIIILALINHEFMFMEVMW
TFSIYLEAVAIMPQLFMLSRTGNAETITAHYLFALGSYRFLYIFNWVYRYTESFFDPIA
VVAGIVQTVLYADFFLYITRVIQSNRQFEMSA
>ERD2_DICDI_Amoebozoa
MNLFSFLGDMHLGSMILILFKIKNDKSCAGVSLKSQILFTIVFTARYLDLFTNVVSLYI
TFMKITYIAVSYYTLHLIARKYKFTYDKDHDTFKIVYLIASCILSLITYDKTTIGIYST
FLEILWTFESIYLESIAILPQLILLQRTGEVEALTSNYIVLLGGYRAFYLFWNIYRITFYN
WSGKIEMLSGLLQTILYADFFYYYAKSRMYGKKLVLPQ
>ERD2_CAEEL_Metazoa
MNLFRFTADVAHAIAIVVLLKKIWKSRSCGISGRSQLLFALVFTRYLDLFTNFFSFYN
TAMKIFYLVASFQTVYLMWAKFKATYDRNNDSFRIEFLVIPSIMILALLINHEFIFMEVMW
TFSIYLEAVAIMPQLFMLSRTGNAETITAHYLFALGSYRFLYILNWVYRYTESFFDPIS
VVAGIVQTVLYADFFLYITRVIQSNRQFEMSA
>ERD2_DROME_Metazoa
MNI FRFAGDLSHVFAIIILLKKIWKTRSCAGISGKSQILFAVVYLYTRYLDLFTTYVSLYN
SVMKVLFLATSGATVYLMYVKFKATYDHNHDSFRIEFLVPCALLSLVINHEFTVMEVLW
TFSIYLESVAAILPQLFLVSRTEAESITSHYLFALGSYRALYLLNWVYRYMVESHYDLIA
IFAGVVQTVLYCDFFLYITKVLKGKKLQLP
>ERD2_ENTHI_Amoebozoa
MVFNLFRIADLVHLLSIYFLLTKIISHKNCIGISLRSQILFFIVWVTRYLDIFYNFYSL
YNTILKIVYLTTSAYTIYILISKRFRTYDKIHDTLNVWYLIVPCIVLAFIFTEDYSITEI
CWTFSIFLEAVAILPQILLRSTGEVENLNSQYIFCLGLYRALYIINWIYRYATEQSYWS
PLTWICGSIQTLTYVEYFYIYIKSRVEGTEKFLVLPY
>ERD2_KLULA_Fungi
MLNVFRIAGDFSHLASIIILIQSITTSNSVDGISLKTQLLYTLVITRYLNLFTKWTSLY
NFLMKIVFISSSVYVIVLMRQKFKNPVAYQDMITRDQFKIKFLIVPCILLGLIFNYRFS
FIQICWSFSLWLESVAAILPQLFMLTKTGKAKQLTSHYIFALGLYRALYIPNWIWRYYTEE
RFDKLSVFTGVIQTLVYSDFFYIYYQKVIKLGGLDLELPQ
>ERD2_PETHY_Viridiplantae
MNI FRLAGDMTHLASVLVLLKIHTIKSCAGVSLKTQELYALVFVTRYLDIFTDFISLYN
TTMKLVFLGSSLSIVWYMRHHKIVRRSYDKDQDTRHFLVLVPCLLLALVINEKFTFKEV
MWTFESIYLEAVAILPQLVLLQRTNRIDNLTGQYIFLLGAYRSFYILNWVYRYFTEPHFVH
WITWIAGLIQTLTYADFFYYYFQSWKNNTKLELPA
>ERD2_PLAF7_Apicomplexa
MNI FRLIGDILHLVSMYILIMKLKSKNCIGISCRMQELYLIVFLCRYIDLFFVFSFYN
TVMKITFILTIAYTIYLIRLKLPISTYNRKVDNFKSEKYLIPCLVLSLLTCKTYNLN
ILWSFSIWLESVAAILPQLVLLQKREVENITSHYVITMGLYRAFYLNWYRYFFDDKPY
INVVGWIGGLIQTLTYIDFFYYFALAKWYGKKLVLPFNGEV
>ERD2_PLAFA_Apicomplexa
MNI FRLIGDILHLVSMYILIMKLKSKNCIGISCRMQELYLIVFLCRYIDLFFVFSFYN
TVMKITFILTIAYTIYLIRLKLPISTYNRKVDNFKSEKYLIPCLVLSLLTCKTYNLN
ILWSFSIWLESVAAILPQLVLLQKREVENITSHYVITMGLYRAFYLNWYRYFFDDKPY
INVVGWIGGLIQTLTYIDFFYYFALAKWYGKKLVLPFNGEV
>ERD2_SCHPO_Fungi
MTFFSALGDMALAAIFLLHMRMKSSTCSGLSLKSHLLFLLVYVTRYLNLFWRYKSLYY
FLMRIVFIASESYICYLMLMTLRPTYDKRLDTRTEYILGGCAVLALIYPTSYTISNILW
TFSIWLESVAAILPQLFMLSGETESLTAHYLFAMCLYRGLYIPHWIYRIAVHKKVIGVA
ILAGIIQTVLYGDAVVYRRTVLQGGKFLRPA
>ERD2_YEAST_Fungi
MNPFRILGDLSHLTSILILIHNIKTTTRYIEGISFKTQTLYALVFITRYLDLLTFHWVSLY
NALMKIFFIVSTAYIVVLLQGSKRNTIAYNEMLMHDTFKIQHLLIGSALMSVFFHHKFT
FLELAWSFSVWLESVAAILPQLYMLSKGGKTRSLTVHYIFAMGLYRALYIPNWIWRYSTED
KKLDKIAFFAGLLQTLTYSDFFYIYYTKVIRGKGFKLPK

Spørgsmål b: Jeg gik til MAFFT serveren på EBI og lavede et multiple alignment af den rettede FASTA fil. Som output format valgte jeg FASTA. Derefter downloadede jeg resultatet ved at højreklikke på “Download Alignment File” i fanen “Alignments”. Dette resultat uploadede jeg til TreeHugger og klikkede på “Submit query”. Så højreklikkede jeg på “Download data in Newick/Phylip format” og gemte filen på min computer. Så startede jeg FigTree og åbnede den gemte fil.

Jeg blev bedt om at bruge slægten *Plasmodium* som *outgroup*. Der er to *Plasmodium* sekvenser i mit datasæt, ERD2_PLAFA og ERD2_PLAF7. Jeg klikker på den gren, der fører til disse to og klikker på “Reroot” øverst i vinduet. Endelig klikker jeg på “Tip Labels” til venstre i vinduet og sætter “Font Size” til 15. Så ser mit træ således ud (efter at det er blevet eksporteret som .png fil):



Spørgsmål c: Dyr (24 sekvenser) og planter (2 sekvenser) er hver især repræsenteret som systematiske grupper, men svampe (3 sekvenser) er ikke, idet gruppen af ERD2_YEAST (*Saccharomyces cerevisiae*) og ERD2_KLULA (*Kluyveromyces lactis*) er tættere beslægtet med dyrene end med ERD2_SCHPO (*Schizosaccharomyces pombe*). Det er en fejl. Der er én fejl mere: ERD2_ENTHI (*Entamoeba histolytica*) og ERD2_DICDI (*Dictyostelium discoideum*) burde høre sammen i en systematisk gruppe (*Amoebozoa*), men det gør de ikke.

Spørgsmål d: Mulighed **2** er den rigtige: Opdelingen i ER lumen protein retaining receptor 1, 2 og 3 er sket tidligt i hvirveldyrenes (*Vertebrata*) udvikling, før de forskellige klasser (såsom benfisk, padder og pattedyr) spaltede ud. Det kan man se af at ERD21-, ERD22- og ERD23-sekvenserne falder i tre adskilte grupper, der hver især indeholder flere klasser af hvirveldyr. Alle tre grupper indeholder både pattedyr (f.eks. HUMAN) og padder (*Xenopus*), mens to af grupperne (ERD22 og ERD23) også indeholder fisk (*Danio rerio*).

Opgave 4 - Forudsigelse af ER lumen proteiner

Spørgsmål a: Det positive testsæt indeholder 107 linjer, og de første 10 linjer ser således ud:

```
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DLEEDDDQKAVRDEL 1
DMEEDDDQKAVKDEL 1
PEPPANSTMGSKEEL 1
PEAQANSTLGPKEEL 1
PEPPANSTMGSKEEL 1
```

Spørgsmål b: I position 13 (den fremhævede position) er der syv D'er, to N'er og ét G. Det giver:

$$f_D = 0.7$$

$$f_N = 0.2$$

$$f_G = 0.1$$

$$g_D = f_D * q(D|D) + f_N * q(D|N) + f_G * q(D|G) = 0.7 * 0.40 + 0.2 * 0.08 + 0.1 * 0.03 = 0.299$$

$$g_E = f_D * q(E|D) + f_N * q(E|N) + f_G * q(E|G) = 0.7 * 0.09 + 0.2 * 0.05 + 0.1 * 0.03 = 0.076$$

$$p_D = (\alpha * f_D + \beta * g_D) / (\alpha + \beta) = (9 * 0.7 + 10 * 0.299) / (9 + 10) = 0.489$$

$$p_E = (\alpha * f_E + \beta * g_E) / (\alpha + \beta) = (9 * 0 + 10 * 0.076) / (9 + 10) = 0.040$$

$$w_D = 2 * \log(p_D / q_D) / \log(2) = 2 * \log(0.489 / 0.054) / \log(2) = 6.358$$

$$w_E = 2 * \log(p_E / q_E) / \log(2) = 2 * \log(0.040 / 0.054) / \log(2) = -0.866$$

Hvis man bruger EasyPred til at kontrollere w-værdierne, får man:

$$w_D = 6.355$$

$$w_E = -0.850$$

Hvilket er tæt nok på, givet at værdierne i tabellen i handoutet er afrundede.

Spørgsmål c: 22649 sekvenser

Søgestreng: `taxonomy:Fungi AND annotation:(type:signal) NOT
annotation:(type:location "endoplasmic reticulum") AND fragment:no`

Spørgsmål d: 333 linjer (107 positive og 226 negative)

Spørgsmål e:

Pearson coefficient: 0.64637

Aroc value: 0.85163

Hvis vores metode var perfekt, ville begge værdier være 1. Hvis vores metode gættede tilfældigt, ville Pearson være 0 og Aroc være 0.5.